Mixture to Mixture: Leveraging Close-Talk Mixtures as Weak-Supervision for Speech Separation

Zhong-Qiu Wang

***Abstract*—We propose *mixture to mixture* (M2M) training, a weakly-supervised neural speech separation algorithm that leverages close-talk mixtures as a weak supervision for training discriminative models to separate far-field mixtures. Our idea is that, for a target speaker, its close-talk mixture has a much higher signal-to-noise ratio (SNR) of the target speaker than any far-field mixtures, and hence could be utilized to design a weak supervision for separation. To realize this, at each training step we feed a far-field mixture to a deep neural network (DNN) to produce an intermediate estimate for each speaker, and, for each of considered close-talk and far-field microphones, we linearly filter the DNN estimates and optimize a loss so that the filtered estimates of all the speakers can sum up to the mixture captured by each of the considered microphones. Evaluation results on a 2-speaker separation task in simulated reverberant conditions show that M2M can effectively leverage close-talk mixtures as a weak supervision for separating far-field mixtures.**

***Index Terms*—Weakly-supervised neural speech separation.**

I. Introduction

DEEP learning has significantly elevated the performance of speech separation [1] thanks to its strong modeling capabilities on human speech, especially since deep clustering [2] and permutation invariant training (PIT) [3] solved the label permutation problem. Modern neural speech separation models [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23] are usually trained on simulated data in a supervised way, where clean speech is synthetically mixed with noise and competing speech in simulated reverberant rooms to generate paired clean and corrupted speech for supervised learning, where the clean speech can provide a *sample-level* supervision for DNN training. The trained models, however, often suffer from mismatches between simulated and real-recorded data, and are known to have severe generalization issues on real-recorded data [24], [25], [26], [27], [28], [29].

One way to address the problem is training models directly on real-recorded mixtures. This however cannot be applied for supervised approaches since it is not possible to annotate the clean speech at each sample. Another way is training unsupervised models on real-recorded mixtures, which usually makes strong assumptions on signal characteristics [27], [28], [29], [30], [31]. However, the performance could be fundamentally limited due to not leveraging any supervision and when the assumptions are not sufficiently satisfied in reality.

Manuscript received 30 January 2024; revised 19 May 2024; accepted 16 June 2024. Date of publication 20 June 2024; date of current version 3 July 2024. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Tsao.

The author is with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: wang.zhongqiu41@gmail.com).

Digital Object Identifier 10.1109/LSP.2024.3417284

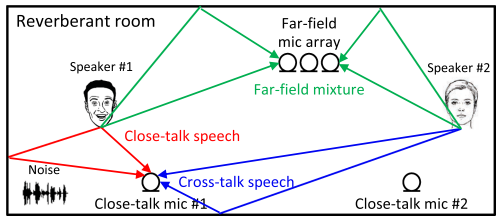

Fig. 1. Illustration of task setup. Each close-talk mixture contains close- and cross-talk speech, and weak noises. Best viewed in color.

While far-field mixtures are recorded in multi-speaker conditions, the close-talk mixture of each speaker is often recorded at the same time by using a close-talk microphone (e.g., in the AMI [32] and CHiME [33] setup). See Fig. 1 for an illustration. The close-talk mixture of each speaker usually has a much higher SNR of the speaker than any far-field mixture. Intuitively, it can be leveraged to train a model to increase the SNR of the speaker in far-field mixtures. In this context, we propose to leverage close-talk mixtures as a weak supervision for separating far-field mixtures. To realize this, we need to solve two major difficulties: (a) close-talk mixtures are often not sufficiently clean, due to the contamination by cross-talk speech [32], [33], [34], [35]; and (b) close-talk mixtures are not time-aligned with far-field mixtures. As a result, close-talk mixtures cannot be naively used as the training targets, and previous studies seldomly exploit them to build separation systems.

To overcome the two difficulties, we propose *mixture to mixture* (M2M) training, where a DNN, taking in far-field mixtures as input, is discriminatively trained to produce an intermediate estimate for each target speaker in a way such that the intermediate estimates for all the speakers can be linearly filtered to recover the close-talk as well as far-field mixtures. Following [29], the linear filters are computed via a neural forward filtering algorithm named forward convolutive prediction (FCP) [36] based on the mixtures and intermediate DNN estimates. We find that this linear filtering procedure can effectively address the above two difficulties. This paper makes two major contributions:

- We are the first seeking a way to leverage close-talk mixtures as a weak supervision for speech separation;
- We propose a novel algorithm named M2M to exploit this weak supervision.

As an initial step, this paper evaluates M2M on a 2-speaker separation task in simulated, reverberant conditions with weak noises. The evaluation results show that M2M can effectively leverage the weak-supervision afforded by close-talk mixtures.

II. Related Work

There are several earlier studies on weakly-supervised separation. In [37], [38], adversarially trained discriminators (in essence, source prior models) are used to encourage separation models to produce

separation results with distributions similar to clean sources. In [39], separation frontends are jointly trained with backend ASR models so that word transcriptions can be used to help frontends learn to separate. In [40], a sound classifier is used to guide separation, by checking whether separated signals can be classified as target sound classes. These approaches need clean sources, human annotations, and other models (e.g., discriminators, ASR models and sound classifiers). Differently, M2M requires paired close-talk and far-field mixtures, which can be readily obtained during data collection by additionally using close-talk microphones, and it does not require other models. On the other hand, close-talk mixtures exploited in M2M can provide a *sample-level* supervision, which is much more fine-grained than source prior models, word transcriptions, and segment-level class labels.

III. Physical Model and Objectives

In a reverberant environment with C speakers (each wearing a close-talk microphone) and a far-field P-microphone array (see Fig. 1), each recorded far-field and closed-talk mixture can be respectively formulated in the short-time Fourier transform (STFT) domain as follows:

$$Y_p(t,f) = \sum\nolimits_{c=1}^{C} X_p(c,t,f) + \varepsilon_p(t,f), \tag{1}$$

$$Y_d(t,f) = \sum\nolimits_{c=1}^{C} X_d(c,t,f) + \varepsilon_d(t,f), \tag{2}$$

where t indexes T frames, f indexes F frequencies, c indexes C speakers, d indexes C close-talk microphones, and p indexes P far-field microphones. At time t and frequency f, $Y_p(t,f)$, $X_p(c,t,f)$ and $\varepsilon_p(t,f)$ in (1) respectively denote the far-field mixture, reverberant image of speaker c, and non-speech signals captured at far-field microphone p. $Y_d(t,f)$, $X_d(c,t,f)$ and $\varepsilon_d(t,f)$ in (2) respectively denote the STFT coefficients of the close-talk mixture, reverberant image of speaker c, and non-speech signals captured at close-talk microphone d at time t and frequency f. In the rest of this paper, we refer to the corresponding spectrograms when dropping indices p, c, d, t or f. In this study, ε is assumed a weak noise.

While speaker c is talking, its close-talk speech $X_d(c)$, with $d = c$, in the close-talk mixture Y_d is typically much stronger than cross-talk speech $X_d(c')$ by any other speaker c' ($\neq c$). By using close-talk mixtures as a weak supervision, we aim at training a DNN that can learn to estimate the reverberant speaker images (i.e., $X_p(c)$ for each speaker c at a reference far-field microphone p), using only far-field mixtures as input.

IV. M2M

Fig. 2 illustrates M2M. The DNN takes in far-field mixtures as input and produces an intermediate estimate $\hat{Z}(c)$ for each speaker c. Each estimate $\hat{Z}(c)$ is then linearly filtered via FCP such that the filtered estimates can be summated to recover each of the close-talk and far-field mixtures. This section describes the DNN setup, loss functions, and FCP filtering.

A. DNN Setup

The DNN is trained to perform complex spectral mapping [10], [11], [12], where the real and imaginary (RI) components of far-field mixtures are stacked as input features for the DNN to predict the RI components of $\hat{Z}(c)$ for each speaker c. The DNN setup is described in Section V and the loss function in Section IV-B.

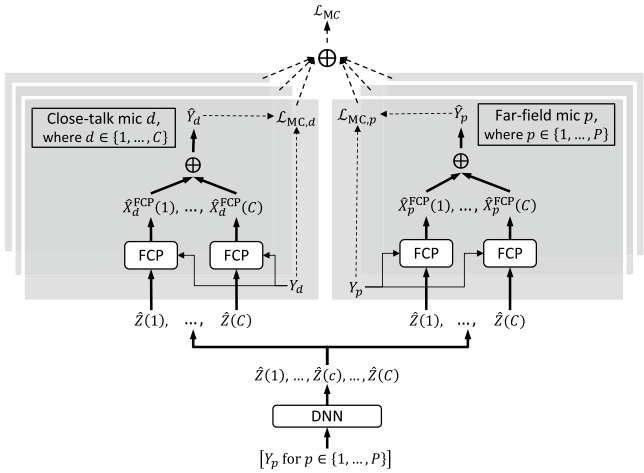

Fig. 2. Illustration of M2M (described in first paragraph of IV).

B. Mixture-Constraint Loss

We propose a mixture-constraint (MC) loss to encourage the DNN to produce an intermediate estimate $\hat{Z}$ that can be utilized to reconstruct the close-talk and far-field mixtures:

$$\mathcal{L}_{\text{MC}} = \sum\nolimits_{d=1}^{C} \mathcal{L}_{\text{MC},d} + \alpha \times \sum\nolimits_{p=1}^{P} \mathcal{L}_{\text{MC},p}, \tag{3}$$

where $\mathcal{L}_{\text{MC},d}$ is the loss at close-talk microphone d, $\mathcal{L}_{\text{MC},p}$ at far-field microphone p, and $\alpha \in \mathbb{R}_{>0}$ a weighting term.

$\mathcal{L}_{\text{MC},d}$ is defined, following the physical model in (2), as

$$\begin{aligned}\mathcal{L}_{\text{MC},d} &= \sum\nolimits_{t,f} \mathcal{F}\left(Y_d(t,f), \hat{Y}_d(t,f)\right) \\ &= \sum\nolimits_{t,f} \mathcal{F}\left(Y_d(t,f), \sum\nolimits_{c=1}^{C} \hat{X}_d^{\text{FCP}}(c,t,f)\right) \\ &= \sum\nolimits_{t,f} \mathcal{F}\left(Y_d(t,f), \sum\nolimits_{c=1}^{C} \hat{\mathbf{g}}_d(c,f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c,t,f)\right),\end{aligned} \tag{4}$$

where $\widetilde{\hat{\mathbf{Z}}}(c,t,f) = [\hat{Z}(c,t-I,f), \ldots, \hat{Z}(c,t+J,f)]^{\mathsf{T}} \in \mathbb{C}^{I+1+J}$ stacks a window of T-F units, $\hat{\mathbf{g}}_d(c,f) \in \mathbb{C}^{I+1+J}$ is a time-invariant FCP filter which will be detailed in Section IV-C, and $\mathcal{F}(\cdot,\cdot)$ is a distance measure to be described later. In (4), the intermediate estimate $\hat{Z}(c)$ of each speaker c is linearly filtered such that (a) the filtering result, $\hat{X}_d^{\text{FCP}}(c,t,f) = \hat{\mathbf{g}}_d(c,f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c,t,f)$, can approximate $X_d(c)$, the cross-talk speech of speaker c captured by close-talk microphone d; and (b) the filtering results of all the speakers can add up to the close-talk mixture Y_d (i.e., $\hat{Y}_d = \sum_{c=1}^{C} \hat{X}_d^{\text{FCP}}(c)$). This way, we can leverage close-talk mixtures as a weak supervision for model training, and the linear filtering procedure can account for the mismatched time-alignment between close-talk and far-field mixtures. Since the model is trained to reconstruct close-talk mixtures based on far-field mixtures, we name the algorithm *mixture to mixture*.

$\mathcal{F}(\cdot,\cdot)$ in (4) computes a loss between the mixture Y_d and reconstructed mixture $\hat{Y}_d$ based on the estimated RI components and their magnitude [29]:

$$\mathcal{F}\left(Y_d(t,f), \hat{Y}_d(t,f)\right) = \frac{\sum_{\mathcal{O}\in\Omega} \left|\mathcal{O}(Y_d(t,f)) - \mathcal{O}(\hat{Y}_d(t,f))\right|}{\sum_{t',f'} \left|Y_d(t',f')\right|}, \tag{5}$$

where $\Omega = \{\mathcal{R}, \mathcal{I}, \mathcal{A}\}$ denotes a set of functions with $\mathcal{R}(\cdot)$ extracting the real part, $\mathcal{I}(\cdot)$ the imaginary part and $\mathcal{A}(\cdot)$ the magnitude of a complex number, $|\cdot|$ computes magnitude, and the denominator balances the losses at different microphones and across training mixtures.

Following (1), $\mathcal{L}_{\text{MC},p}$ is similarly defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{MC},p} &= \sum\nolimits_{t,f} \mathcal{F}\left(Y_p(t,f), \hat{Y}_p(t,f)\right) \\ &= \sum\nolimits_{t,f} \mathcal{F}\left(Y_p(t,f), \sum\nolimits_{c=1}^{C} \hat{X}_p^{\text{FCP}}(c,t,f)\right) \\ &= \sum\nolimits_{t,f} \mathcal{F}\left(Y_p(t,f), \sum\nolimits_{c=1}^{C} \hat{\mathbf{g}}_p(c,f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c,t,f)\right), \end{aligned} \tag{6}$$

where $\widetilde{\hat{\mathbf{Z}}}(c,t,f) = [\hat{Z}(c,t-M,f), \ldots, \hat{Z}(c,t+N,f)]^{\mathsf{T}} \in \mathbb{C}^{M+1+N}$ stacks a window of T-F units and $\hat{\mathbf{g}}_p(c,f) \in \mathbb{C}^{M+1+N}$ is a time-invariant FCP filter to be described later.

We can configure the filter taps, I, J, M and N, differently for close-talk and far-field microphones, considering that the microphones form a distributed rather than compact array.

C. FCP for Relative RIR Estimation

To compute $\mathcal{L}_{\text{MC}}$, the linear filters need to be first computed. Each filter can be interpreted as the relative transfer function (RTF) relating the intermediate DNN estimate of a speaker to its reverberant image captured by another microphone. Following [29], we employ FCP [36] to estimate the RTFs.

Assuming that speakers do not move within each utterance, we estimate RTFs by solving the following problem:

$$\hat{\mathbf{g}}_r(c,f) = \underset{\mathbf{g}_r(c,f)}{\operatorname{argmin}} \sum_t \frac{\left|Y_r(t,f) - \mathbf{g}_r(c,f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c,t,f)\right|^2}{\hat{\lambda}_r(c,t,f)}, \tag{7}$$

where the subscript r indexes the P far-field and C close-talk microphones, and $\mathbf{g}_r(c,f)$ and $\widetilde{\hat{\mathbf{Z}}}(c,t,f)$ are defined in the text below (4) and (6). $\hat{\lambda}$ is a weighting term balancing the importance of each T-F unit. For each close-talk microphone d and speaker c, it is defined, following [36], as

$$\hat{\lambda}_d(c,t,f) = \xi \times \max(|Y_d|^2) + |Y_d(t,f)|^2, \tag{8}$$

where ξ (set to 10^{-4}) floors the weighting term and $\max(\cdot)$ extracts the maximum value of a power spectrogram; and for each far-field microphone p and speaker c, it is defined as

$$\hat{\lambda}_p(c,t,f) = \xi \times \max(Q) + Q(t,f), \tag{9}$$

where $Q = \frac{1}{P}\sum_{p=1}^{P} |Y_p|^2$ averages the power spectrograms of far-field mixtures. Notice that $\hat{\lambda}$ is computed differently for different microphones, as the energy level of each speaker is different at close-talk and far-field microphones. Notice that (7) is a quadratic problem, where a closed-form solution can be readily computed. We then plug $\hat{\mathbf{g}}_r(c,f)$ into (4) and (6) to compute the loss, and train the DNN.

Although, in (7), $\hat{Z}(c)$ is linearly filtered to approximate Y_r, previous studies [29], [36] have suggested that the filtering result $\hat{\mathbf{g}}_r(c,f)^{\mathsf{H}}\widetilde{\hat{\mathbf{Z}}}(c,t,f)$ would approximate the speaker image $X_r(c,t,f)$, when $\hat{Z}(c)$ gets sufficiently accurate during training so that $\hat{Z}(c)$ becomes little correlated with sources other than c (see detailed derivations in Appendix C of [29]). The estimated speaker image is named *FCP-estimated image:*

$$\hat{X}_r^{\text{FCP}}(c,t,f) = \hat{\mathbf{g}}_r(c,f)^{\mathsf{H}}\, \widetilde{\hat{\mathbf{Z}}}(c,t,f). \tag{10}$$

The FCP-estimated images of all the speakers can be hence summated to reconstruct Y_r in (4) and (6).

At run time, we use the FCP-estimated image $\hat{X}_p^{\text{FCP}}(c)$ as the prediction for each speaker c at a reference far-field microphone p. We use the time-domain signal of the clean far-field image, $X_p(c)$, as the reference signal for evaluation.

TABLE I
QUALITY OF CLOSE-TALK MIXTURES (REF: CLOSE-TALK SPEECH)

Dataset	SI-SDR (dB)↑	SDR (dB)↑	PESQ↑	eSTOI↑
SMS-WSJ-FF-CT	14.7	14.7	2.92	0.875

D. Relations to, and Differences From, UNSSOR

M2M is motivated by a recent algorithm named UNSSOR [29], an unsupervised neural speech separation algorithm designed for separating far-field mixtures. UNSSOR is trained to optimize a loss similar to (6), by leveraging the mixture signal at each microphone as a constraint to regularize DNN-estimated speaker images, and it can be successfully trained if the mixtures for training are over-determined (i.e., more microphones than sources) [29]. The major novelty of M2M is adapting UNSSOR for weakly-supervised separation by defining the MC loss not only on far-field microphones, but also on close-talk microphones to leverage the weak supervision afforded by close-talk mixtures to obtain better separation than UNSSOR, which is unsupervised. In M2M, there are C speakers, and P far-field and C close-talk microphones for loss computation. The over-determined condition is hence naturally satisfied (i.e., $P + C > C$). With that being said, only using the MC loss on close-talk mixtures (i.e., the first term in (3)) for training M2M would not lead to separation of speakers. This is because, as is suggested in UNSSOR [29], the number of close-talk mixtures used for loss computation is not larger than the number of sources, and there would be an infinite number of DNN-estimated speaker images that can minimize the MC loss. The second term in (3) can help narrow down the infinite solutions to target speaker images.

V. EXPERIMENTAL SETUP

Since there are no earlier studies leveraging close-talk mixtures as a weak supervision for separation, to validate M2M we propose a simulated dataset so that clean reference signals can be used for evaluation. Building upon the SMS-WSJ corpus [41], which only has far-field (FF) mixtures, we simulate SMS-WSJ-FF-CT, by adding close-talk (CT) mixtures.

SMS-WSJ [41] is a popular corpus for 2-speaker separation in reverberant conditions. It has 33,561 (∼87.4 h), 982 (∼2.5 h) and 1,332 (∼3.4 h) 2-speaker mixtures respectively for training, validation and testing. The clean speech is sampled from the WSJ0 and WSJ1 corpus. The simulated far-field array has 6 microphones uniformly placed on a circle with a diameter of 20 cm. For each mixture, the speaker-to-array distance is drawn from the range $[1.0, 2.0]$ m, and the reverberation time (T60) from $[0.2, 0.5]$ s. A white noise is added, at an energy level between the summation of the reverberant speech and the noise, sampled from the range $[20, 30]$ dB. **SMS-WSJ-FF-CT** is created by adding a close-talk microphone for each speaker in each SMS-WSJ mixture. The distance from each speaker to its close-talk microphone is uniformly sampled from the range $[10, 30]$ cm. All the other setup for simulation remains the same. This way, we can simulate the close-talk mixture of each speaker, and the far-field mixtures are exactly the same as the existing ones in SMS-WSJ. The sampling rate is 8 kHz.

For STFT, the window size is 32 ms and hop size 8 ms. TF-GridNet [18], which has shown strong performance in major supervised speech separation benchmarks, is used as the DNN architecture. Using the symbols defined in Table I of [18], we set its hyper-parameters to $D = 96$, $B = 4$, $I = 2$, $J = 2$, $H = 192$, $L = 4$ and $E = 4$ (please do not confuse these symbols with the ones in this paper). We train it on 4-second segments using a batch size of 4. The first far-field microphone is designated as the reference microphone. We consider 6-channel

TABLE II
SEPARATION RESULTS ON SMS-WSJ-FF-CT ($P = 6$) (REFERENCE: SPEAKER IMAGES AT FAR-FIELD REFERENCE MIC)

Row	Type	Systems	$I/J/M/N$	α	SI-SDR (dB)↑	SDR (dB)↑	PESQ↑	eSTOI↑
0	-	Mixture	-	-	−0.0	0.1	1.87	0.603
1a	weakly-sup.	M2M	19/1/19/1	1.0	16.9	17.9	3.85	0.931
1b	weakly-sup.	M2M	19/2/19/1	1.0	15.8	16.7	3.76	0.916
1c	weakly-sup.	M2M	19/3/19/1	1.0	16.3	17.2	3.79	0.924
1d	weakly-sup.	M2M	19/4/19/1	1.0	15.9	16.8	3.74	0.917
1e	weakly-sup.	M2M	19/5/19/1	1.0	15.8	16.7	3.75	0.914
2a	weakly-sup.	M2M	19/0/19/0	1.0	16.0	16.9	3.77	0.918
2b	weakly-sup.	M2M	19/2/19/2	1.0	15.8	16.8	3.74	0.914
2c	weakly-sup.	M2M	19/3/19/3	1.0	16.1	17.1	3.77	0.921
4a	Unsupervised	UNSSOR [29]	-	-	14.7	15.6	3.44	0.886
4b	Supervised	PIT [3]	-	-	18.9	19.4	4.06	0.950

TABLE III
SEPARATION RESULTS ON SMS-WSJ-FF-CT ($P = 1$) (REFERENCE: SPEAKER IMAGES AT FAR-FIELD REFERENCE MIC)

Row	Type	Systems	$I/J/M/N$	α	SI-SDR (dB)↑	SDR (dB)↑	PESQ↑	eSTOI↑
0	-	Mixture	-	-	−0.0	0.1	1.87	0.603
1	weakly-sup.	M2M	19/1/19/1	1.0	−2.6	−1.3	1.64	0.479
2a	weakly-sup.	M2M	19/1/19/1	1/5	12.0	12.9	3.41	0.857
2b	weakly-sup.	M2M	19/1/19/1	1/6	11.8	12.7	3.40	0.853
2c	weakly-sup.	M2M	19/1/19/1	1/7	12.7	13.7	3.50	0.872
2d	weakly-sup.	M2M	19/1/19/1	1/8	12.1	13.0	3.45	0.859
3	Supervised	PIT [3]	-	-	13.7	14.1	3.61	0.884

separation, where all the 6 far-field microphones are used as input to M2M, and 1-channel separation, where only the reference microphone signal can be used as input. The evaluation metrics include signal-to-distortion ratio (SDR) [42], scale-invariant SDR (SI-SDR) [43], perceptual evaluation of speech quality (PESQ) [44], and extended short-time objective intelligibility (eSTOI) [45].

For comparison, we consider an unsupervised neural speech separation algorithm named UNSSOR [29], which is trained on far-field mixtures without using any supervision but also by optimizing a loss defined between linearly-filtered DNN estimates and observed mixtures. In addition, we provide the results of PIT [3], trained in a supervised way assuming the availability of clean speaker images at far-field microphones, by using a loss defined, similarly to (5), on the predicted real, imaginary and magnitude components. Both baselines use the same TF-GridNet architecture and training setup as M2M, and their performance can be respectively viewed as the lower- and upper-bound performance of M2M.

VI. EVALUATION RESULTS

Table I presents the scores of close-talk mixtures, which are computed by using the close-talk speech of each speaker as reference and close-talk mixture as estimate. We can see that the close-talk mixtures are not sufficiently clean (e.g., only 14.7 dB in SI-SDR), due to the contamination by cross-talk speech, but the SNR of the target speaker is reasonably high.

Table II reports the results of M2M when there are $P = 6$ far-field microphones. The reference signals for metric computation are the speaker images captured by the far-field reference microphone. In row 1a-1e and 2a-2c, we tune the filter taps I, J, M and N, and observe that the setup in 1a leads to the best separation. Only one future tap (i.e., $J = 1$ and $N = 1$) is used in row 1a, and using more future taps are not helpful, likely because sound would travel $2.72 = 340 \times 0.008$ meters in 8 ms (equal to the hop size of our system) if its speed in air is 340 m/s, and this distance is already larger than the aperture size formed by the simulated close-talk and far-field microphones. Compared to unsupervised UNSSOR in 4a, M2M in 1a produces clearly better separation; and compared with fully-supervised PIT in 4b, M2M in 1a shows competitive performance. These results indicate that M2M can effectively leverage the weak supervision afforded by close-talk mixtures

Table III reports the results when $P = 1$, where M2M only takes in the far-field mixture signal at the reference microphone as input and is trained to reconstruct the input mixture and close-talk mixtures. When the weight α in (3) is 1.0, the DNN could just copy its input as the output (e.g., $\hat{Z}(c) = Y_1$) to optimize the loss on the far-field mixture (i.e., the second term in (3)) to zero, causing the loss on close-talk mixtures not optimized well. To avoid this, we apply a smaller weight α to the loss on the far-field mixture so that the DNN can focus on reconstructing close-talk mixtures. From row 1 and 2a-2 d of Table III, we can see that this strategy works, and M2M obtains competitive results compared to monaural supervised PIT in row 3.

In our experiments, we observe that, even if FCP is performed in each frequency independently from the others, M2M does not suffer from the frequency permutation problem [46], [47], which needs to be carefully dealt with in UNSSOR [29] and in many frequency-domain blind source separation algorithms [47]. This is possibly because each close-talk mixture has a high SNR of the target speaker, which can give a hint to M2M regarding what the target source is across all the frequencies of each output spectrogram.

A sound demo based on the experiments is provided in this link https://zqwang7.github.io/demos/M2M_demo/index.html.

VII. CONCLUSION

We have proposed M2M, which leverages close-talk mixtures as a weak supervision for training neural speech separation models to separate far-field mixtures. Evaluation results on 2-speaker separation in simulated conditions show the effectiveness of M2M. Future research will modify and evaluate M2M on real-recorded far-field and close-talk mixtures.

In closing, the key scientific contribution of this paper, we emphasize, is a novel methodology that directly trains neural source separation models based on paired mixtures, where the higher-SNR mixture can serve as a weak supervision for separating the lower-SNR mixture. This concept of *mixture-to-mixture training*, we believe, would motivate the design of many algorithms in future research in neural source separation.

REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
[2] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.
[3] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.
[4] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[6] M. MacIejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 696–700.
[7] T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, "The cone of silence: Speech separation by localization," in *Proc. Annu. Conf. Neural. Inf. Process. Syst.*, 2020, pp. 20925–20938.
[8] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7121–7132.
[9] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.
[10] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.
[11] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.
[12] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, 2021.
[13] J. Zhang, C. Zorila, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6389–6393.
[14] Z. Zhang et al., "Multi-channel multi-frame ADL-MVDR for target speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3526–3540, 2021.
[15] R. Gu, S. X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Process. Lett.*, vol. 28, pp. 1370–1374, 2021.
[16] Y. Luo, "A time-domain generalized wiener filter for multi-channel speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3008–3019, 2022.
[17] T. Yoshioka et al., "VarArray: Array-geometry-agnostic continuous speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6027–6031.
[18] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.
[19] J. Rixen and M. Renz, "SFSRNet: Super-resolution for single-channel audio source separation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 11220–11228.
[20] E. Tzinis, G. Wichern, A. Subramanian, P. Smaragdis, and J. L. Roux, "Heterogeneous target speech separation," in *Proc. Interspeech*, 2022, pp. 1796–1800.
[21] K. Patterson, K. Wilson, S. Wisdom, and J. R. Hershey, "Distance-based sound separation," in *Proc. Interspeech*, 2022, pp. 901–905.
[22] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Cernocky, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Process. Mag.*, vol. 40, no. 3, pp. 8–29, May 2023.
[23] S. R. Chetupalli and E. A. Habets, "Speaker counting and separation from single-channel noisy mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1681–1692, 2023.
[24] S. Cornell et al., "The CHiME-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios," in *Proc. CHiME*, 2023.
[25] R. Aralikatti, C. Boeddeker, G. Wichern, A. Subramanian, and J. L. Roux, "Reverberation as supervision for speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[26] S. Leglaive et al., "The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement," in *Proc. CHiME*, 2023.
[27] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixture invariant training," in *Proc. Annu. Conf. Neural. Inf. Process. Syst.*, 2020, pp. 3846–3857.
[28] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, "RemixIT: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1329–1341, Oct. 2022.
[29] Z.-Q. Wang and S. Watanabe, "UNSSOR: Unsupervised neural speech separation by leveraging over-determined training mixtures," in *Proc. Annu. Conf. Neural. Inf. Process. Syst.*, 2023, pp. 34021–34042.
[30] Y. Bando, Y. Masuyama, A. A. Nugraha, and K. Yoshii, "Neural fast full-rank spatial covariance analysis for blind source separation," in *Proc. 31st Eur. Signal Process. Conf.*, 2023, pp. 51–55.
[31] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech," in *Proc. Eur. Signal Process. Conf.*, 2021, pp. 436–440.
[32] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," in *Proc. Mach. Learn. Multimodal Interaction*, 2006, pp. 28–39.
[33] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. Interspeech*, 2018, pp. 1561–1565.
[34] S. Watanabe et al., "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," 2020, *arXiv:2004.09249*.
[35] F. Yu et al., "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6167–6171.
[36] Z.-Q. Wang, G. Wichern, and J. L. Roux, "Convolutive prediction for monaural speech dereverberation and noisy-reverberant speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3476–3490, 2021.
[37] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2391–2395.
[38] N. Zhang, J. Yan, and Y. Zhou, "Weakly supervised audio source separation via spectrum energy preserved Wasserstein learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 4574–4580.
[39] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "MIMO-SPEECH: End-to-end multi-channel multi-speaker speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 237–244.
[40] F. Pishdadian, G. Wichern, and J. L. Roux, "Finding strength in weakness: Learning to separate sounds with weak supervision," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2386–2399, 2020.
[41] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," 2019, *arXiv:1910.13934*.
[42] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[43] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 626–630.
[44] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.
[45] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[46] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
[47] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal Info. Process.*, vol. 8, pp. 1–14, 2019.